USING ELICITED ORAL RESPONSE TESTING TO DETERMINE THE NEED FOR AN INTERPRETER

Dr. William G. Eggington
and
*Dr. Troy Cox**

As increasing numbers of second-language speakers interact with the U.S. legal system, it is often difficult to determine if their speaking and listening abilities are sufficiently developed to accurately comprehend and communicate information relevant to the case at hand. The miscommunication that results can lead to wrongful convictions and costly defenses. This paper examines the potential for using Elicited Oral Response (EOR) tests, in which an examinee hears a sentence and repeats what is said, as a means of determining if a person's language ability is sufficient to function within the legal system without the need for an interpreter. The article reports on two cases where EOR was successfully used as part of the language assessment battery in legal settings.

TABLE OF CONTENTS

* Dr. William G. Eggington is Professor and Department Chair, Linguistics and English Language Department, Brigham Young University, Provo, Utah. Dr. Eggington is a general forensic linguist who also specializes in language and law issues involving linguistic minorities. Dr. Troy Cox is Teaching Assessment Coordinator at the English Language Center, Brigham Young University, Provo, Utah. The authors thank Dr. Deryle Lonsdale and Dr. C. Ray Graham for their pioneering work in Elicited Oral Imitation and Automatic Speech Recognition. We also thank Adam Prestidge, currently a law student at William & Mary Law School, for boldly taking first steps that led to this work.

Introduction

Hamza,[1] an African immigrant, travels on a domestic flight in the United States. He harmlessly flirts with his young seatmate, until she becomes angry with him and moves. When he exits the plane, the police approach him and request an interview. He quickly realizes that something serious is happening and he is being interrogated about his conversation with the young lady. He does not understand what is occurring, but because, in his culture, compliance with authority is seen as a practical and civic duty, he agrees with many of the police's statements. Unfortunately, his English language ability is too poor for him to realize that, in the end, he has unwittingly confessed to sexual battery. He is subsequently arrested. This case will be addressed in more detail later, but the underlying issues are relevant to many contexts involving immigrants with limited English speaking ability.

As immigrants continue to enter the United States in large numbers, miscommunication problems between limited English speakers and the U.S. legal system multiply.[2] According to the Pew Hispanic Center, a record 40.4 million immigrants lived in the United States in 2011.[3] This figure represents, by far, the largest immigrant population of any nation with Russia's 12.3 million immigrants coming in a distant second.[4] Approximately one quarter of the non-U.S.-born population originate from one country, Mexico,[5] with a majority of the remaining immigrant population coming from other areas of Central and South America including Spanish-speaking areas of the Caribbean.[6] In fact, the Spanish language is spoken by 60% of the total foreign-born population in the United States who speak a language other than English at home. Consequently, although the following discussion will initially focus on language proficiency testing in legal contexts for the general non-English-speaking U.S. population, it is obvious that a significant portion of the interaction between the U.S. legal system and the immigrant population involves people from Latino backgrounds.

Sometimes, officials who represent the U.S. legal system such as police officers and other law enforcement personnel overestimate nonnative speakers' fluency due to their limited experience interacting with nonnative speakers outside of law enforcement or legal settings. This inexperience can result in officials failing to notice cues that indicate a lack of full comprehension.

---

[1] The name has been changed to protect the subject's privacy.

[2] *See* Alan M. Maxwell, Hon. Lynn W. Davis, Adam Prestidge and William G. Eggington *Finding Justice in Translation: American Jurisprudence Affecting Due Process for People with Limited English Proficiency Together with Practical Solutions*, 14 Harv. Latino L. Rev. 117 (2011).

[3] Pew Hispanic Ctr., A Nation of Immigrants: A Portrait of the 40 Million, Including 11 Million Unauthorized (Jan. 29, 2013), *available at* http://www.pewhispanic.org/2013/01/29/a-nation-of-immigrants/.

[4] *Id.*

[5] Maxwell et al., *supra* note 2, at 119.

[6] *Id.*

Officials might also fail to perform comprehension checks during the interview, such as asking interviewees to explain what they just heard, or rephrasing questions to verify that the original question was understood. Unfortunately, limited English speakers often employ coping strategies—such as pretending to understand while they try to get more language input in order to decode exactly what is happening—especially in situations where there is a significant power differential. This strategy of feigned comprehension works well in mundane, non-threatening language learning contexts, but can have dire consequences in high-stakes contexts such as an interrogation setting or an interview with a judge.

Those at greatest risk of appearing more fluent than they actually are include language learners who already possess some English-language ability, but are not proficient enough to fully comprehend the language. If an interviewee, suspect, or defendant does not understand any English, the legal system has no choice but to provide an interpreter. However, if a limited English speaker has a repertoire of stock phrases and learned material, as many do, she can appear quite fluent even if her actual ability is inadequate to follow a conversation. On the American Council of the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview Scale (OPI), generally accepted as a proficiency rating standard, this type of limited English speaker would be rated as "Novice-High."[7] At this level, speakers can be somewhat creative with English, but in the course of conversation with a native English speaker who uses normal speech rates[8] and vocabulary, they fail to fully comprehend essential information. They are frequently unable to reply appropriately, often experiencing communication breakdowns requiring sympathetic listeners who are willing to speak more slowly and are experienced in "repairing" communicative problems.[9] A situation where a limited English speaker interacting with a law official who is inexperienced in dealing with limited English speakers, and who is unsympathetic to the difficulties encountered by language learners, will likely lead to serious miscommunication. In addition, these interactions often occur within high-anxiety, high-power-differential settings, with the limited English speaker feigning comprehension as well as using stock, formulaic expressions. This situation often contributes to dangerous outcomes, including law-enforcement officials confusing lack of comprehension for assent.

These communication breakdowns are consequential. They can lead to unjust incarcerations, unnecessary legal expenses, and time away from work and family. With respect to issues involving social cohesion, continuing

---

[7] American Council on the Teaching of Foreign Languages, ACTFL Proficiency Guidelines 2012 (2012) [hereinafter ACTFL Proficiency Guidelines 2012].

[8] Speech rates are conventionally measured by syllables per second (SPS). A range of between 3.8 SPS and 5 SPS is considered normal. *See* Roger Griffiths, *Speech Rate and Listening Comprehension: Further Evidence of the Relationship*, 26 TESOL Q., no. 2, 1992 at 385.

[9] ACTFL Proficiency Guidelines 2012, *supra* note 7.

communicative difficulties can contribute to individuals, families, and immigrant groups developing a significant lack of trust in the U.S. legal system.

In addition, prosecuting the wrong person creates an opportunity cost in the time police could have spent pursuing the actual criminal. When and if police find the right person, evidence and confessions acquired through these initial unintentional, *feigned* confessions allow the defense to file motions to suppress evidence.

A solution to this problem lies in the legal system's potential ability to determine the English language proficiency of a nonnative English speaker prior to engaging in meaningful conversation. Unfortunately, the established methods of assessing oral language ability are problematic even for expert language educators.

The most common method for assessing speaking proficiency has consisted of one-on-one, face-to-face interviews.[10] To mitigate the tendency of interviews to follow their own idiosyncratic patterns of scoring, a second rater is required to verify the original rating, thus increasing the cost.[11] The use of computers for assessing language proficiency has been explored as a potential means of decreasing the time and labor needed to obtain and assess ratable speech samples, but most of the research in this area has been conducted in controlled academic settings.[12] If computerized testing were implemented in legal settings, however, scoring would still require trained personnel, which could be time-consuming.[13]

One alternative that may reduce costs and the necessity of trained language personnel is the use of Elicited Oral Response ("EOR") tests. EOR tests have examinees listen to specific phrases in a foreign language, typically sentence length, and then repeat what they hear. When sufficiently long utterances are used, the examinee must process the language, including grammar, vocabulary, and other linguistic features, to understand the meaning and then reconstruct the sentence to repeat it. The rationale is that examinees cannot process language that is beyond their proficiency level.[14] The difficulty of EOR items can be varied by modifying the factors needed for comprehension: that is, the number of syllables in the sentence, or the grammatical and lexical complexity of the sentences.[15] These tests can also be administered through a variety of means ranging from face-to-face interactions to computerized testing. The ease and cost-effectiveness of using EOR

---

[10] Sari Luoma, Assessing Speaking (2004).

[11] Glenn Fulcher, Testing Second Language Speaking (2003).

[12] Carol A. Chapelle & Dan Douglas, Assessing Language through Computer Technology (2006).

[13] H. Douglas Brown, Language Assessment: Principles and Classroom Practices (1st ed. 2003).

[14] Thora Vinther, *Elicited Imitation: A Brief Overview*, 12 Int'l J. of Applied Linguistics, no. 1, 2002, at 54.

[15] Robert Bley-Vroman & Craig Chaudron, *Elicited Imitation as a Measure of Second-Language Competence*, *in* Res. Methodology in Second-Language Acquisition 245 (Elaine E. Tarone et al. eds., 1994).

for language testing in legal settings form the motivation for the study described below.

I. Research Purpose and Questions

In order to examine the potential of using EOR to assess speaking ability and to determine if an interpreter is needed, the authors conducted a review of the current use of EOR for language assessment in educational settings. To examine how the EOR test functions in legal settings, the authors conducted two case studies which will be described below. The studies addressed the following research questions:

- How did the results of the EOR test compare with other methods used to assess speaking ability?
- What other innovations could make the EOR test a tool that could be easily administered by those without language training, particularly in legal contexts?

We will first provide an overview of language testing practices and apply these practices to challenges faced within the legal sphere. This overview will be followed by a discussion of Elicited Oral Response (EOR) testing procedures using Automatic Speech Recognition (ASR) as a scoring device as a solution to some of the previously mentioned challenges. We will then describe two case studies where a combination of EOR and ASR were used as a valid, reliable, and practical means of evaluating English language proficiency in legal contexts. The article concludes with recommendations for further development of EOR/ASR approaches for legal purposes.

II. Basic Principles of Speaking Assessments

Before an evaluation on the merits of any type of assessment can be made, it is important to understand the basic principles of assessment. There is no such thing as a perfect test that will work well in all circumstances. Every test has strengths and weaknesses and must be examined in that light. Furthermore, while assessments frequently occur, often they are informal assessments that rely on personal subjective opinion rather than objective measurement.[16] As indicated below, three key factors need to be taken into consideration in determining whether a test will function in the environment in which it is intended: reliability, validity, and practicality.[17]

---

[16] Cathleen G. Spinelli, *Addressing the Issue of Cultural and Linguistic Diversity and Assessment: Informal Evaluation Measures for English Language Learners,* Reading and Writing Q., Jan. 2008, at 101.

[17] Jorge Cubillos, *Computer-mediated Oral Proficiency Assessments: Validity, Reliability and Washback,* 6 Int'l J. of Tech., Knowledge and Soc'y, no. 6, at 85.

Reliability refers to consistency in measurement.[18] This construct includes (1) equivalent forms reliability (meaning, the degree to which examinees would get a similar score on an equivalent form of the test); (2) test/retest reliability (that is, the degree to which examinees get similar scores when taking the same test); and (3) inter-rater reliability (that is, the degree to which different raters would give similar scores to the same examinee).[19] When reliability is high, examinees participating in an assessment receive similar scores regardless of the exact content of the test, the day and time they took it, or the raters scoring the test.[20]

Validity refers to whether or not a test measures what it is supposed to measure.[21] Specific evidence of validity is revealed through *construct validity*, or the degree to which test scores can be used to make inferences about specific characteristics or features of an examinee's language ability.[22] For a speaking test to have construct validity, it is important that the examinee actually speak. It would be difficult to claim a test measures speaking ability if examinees simply answered multiple-choice listening questions.

Finally, *practicality* determines whether a test should or should not be administered.[23] To determine if a test is impractical, we must consider the personnel needed, resource requirements, and cost of administration. A test might meet all reliability and validity requirements, but fail the practicality criterion. In fact, practicality is often the overriding factor in deciding whether a test will be used.[24]

III. Oral Proficiency Interviews

Using reliability, validity, and practicality as the framework for evaluating language testing in legal contexts, it is instructional to review one of the most commonly used and highly regarded assessments of speaking ability: the American Council of the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview (OPI).[25] The OPI is a structured interview between an examinee and a certified tester that lasts between fifteen and thirty minutes.[26] For quality purposes, the interview is recorded and subsequently double rated. If there is a discrepancy between the two ratings, additional certified testers resolve the dispute. To become a certified tester, an individual needs

---

[18] Lyle F. Bachman & Adrian S. Palmer, Language Testing in Practice (Oxford University Press, 4th ed. 1996).

[19] *Id.*

[20] Brown, *supra* note 13.

[21] Bachman & Palmer, *supra* note 18.

[22] *Id. See also* Grant Henning, *Oral Proficiency Testing: Comparative Validities of Interview, Imitation, and Completion Methods*, 33 Language Learning, no. 3, 1983, at 315.

[23] Bachman & Palmer, *supra* note 18.

[24] *Id.*

[25] Fulcher, *supra* note 11.

[26] *Id.*

to undergo time-intensive training and supervised practical experience that can take over sixty hours.[27]

Applying the above framework, we can see that there are both strengths and weaknesses in using the OPI. Reliability is improved by the extensive training and multiple ratings of certified testers.[28] One weakness, though, is that the OPI allows for fewer independent samples of speech to rate than would be possible by other testing methods. For example, if the interview happens to be on a topic about which the examinee knows very little, she would score lower than if it were on a topic in which she was well versed. The more different, independent samples that exist, the greater the reliability will be. With regards to the validity of the OPI, since the exam is an oral interview, it could be argued that the score reflects the construct of speaking, and thus would be considered to have high construct validity. However, if specific structures or types of speech need to be assessed, it can be difficult for the interviewer to elicit those forms, and it may be easy for an examinee to avoid them. Thus, the test might have weak content validity. The practicality of having certified raters available to administer assessments when needed in a legal setting can be cost prohibitive. It is expensive to train interviewers, and the one-on-one nature of an interview procedure introduces time constraints that make this kind of testing difficult to use in a real world setting. In legal contexts, even with these faults, using the OPI would still be better than providing no formal language assessment.

IV. Potential of Elicited Oral Response Testing

EOR testing is based on a psycholinguistic research technique often referred to as elicited imitation.[29] With this technique, examinees listen to an utterance, typically a sentence, and then repeat what they hear.[30] For example, examinees might hear a sentence such as, "I bought the jacket last year when it was on sale," delivered at normal speed. They are then required to repeat the sentence. A complete test will provide a range of simple and complex sentence types.

Use of EOR as a speaking assessment is based on two concepts. First, all second language learners develop an interlanguage, which is a "learner language that is implicitly embedded in their cognitive framework."[31] The

---

[27] Kathryn Buck, Heidi Byrnes & Irene Thompson, The ACTFL Oral Proficiency Interview Tester Training Manual (1989).

[28] *Id.*

[29] Alistair Van Moere, *A Psycholinguistic Approach to Oral Language Assessment*, 29 Language Testing, no. 3, 2012, at 325.

[30] This article uses the term EOR for a number of reasons. First, the article aims to differentiate its use as a testing method rather than a research protocol. Second, the article aims to emphasize that more than mere rote imitation and repetition occurs in this exercise. Third, some of the concerns about the use of elicited imitation as a research protocol are nonissues when it is viewed as a testing procedure.

[31] Larry Selinker, *Interlanguage*, 10 Int'l Rev. of Applied Linguistics in Language Teaching, no. 1-4, 1972, at 209.

structures of this interlanguage are influenced by many factors, including the person's native language as well as some universal stages of grammar acquisition through which all language learners pass regardless of their native language.[32] This interlanguage system is similar to the stages children go through when acquiring their native language. For example, children often create phrases that are grammatically incorrect, while yet still systematic, within children's developing linguistic competences.[33] To illustrate, "I eated cookies yesterday," is an overgeneralization of the regular past tense marker *-ed*, to the irregular verb *eat*. Similarly, second language learners misapply these types of analogical patterns as they attempt to acquire a new language.

The second fundamental concept of EOR relevant to its use in testing situations is that short-term, or working, memory has limits.[34] The amount of information that can be stored in working memory is directly related to the ability of the examinee to access long-term memory and the capacity of her interlanguage skills to deconstruct the content into usable units of information. An examinee listening to the utterance to be repeated must understand the content—including the vocabulary used—prior to reconstructing the sentence. The degree to which the examinee can reproduce the sentence depends on the interaction between working memory and long-term memory. Thus, the ability to repeat longer sentences depends on the examinees' ability and knowledge of the language, not just their ability to parrot what is heard.

The capacity of working memory has traditionally been viewed as ranging from five to nine individual pieces of information, where an individual piece of information is defined as, for example, one number in a phone number sequence;[35] however, more recent research indicates that the amount of separate pieces of information an individual can process and immediately recall might be closer to four items.[36] In practical terms, relying on working memory, many adults can repeat a seven-digit phone number, but will falter at repeating a fifteen-digit credit card number. Interestingly, the amount of information one can process and recall increases if the information is "chunked" (using linguistic terminology) into meaningful groups. For example, the string FBICIANSADOD would be difficult for most people to process and remember if each letter was perceived as one piece of information; however, those who are familiar with U.S. government agencies can chunk the letters into meaningful units, as the acronyms FBI, CIA, NSA, and DOD. An individual who has the ability to chunk the letters would much more likely be able to repeat the sequence correctly. Similarly, as one becomes

---

[32] *See* ROD ELLIS & GARY BARKHUIZEN, ANALYSING LEARNER LANGUAGE (2005).

[33] *Id.*

[34] Nelson Cowan, *The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity*, 24 THE BEHAVIORAL AND BRAIN SCI., no. 1, 2000, at 87, 114, *available at* http://www.pri.kyoto-u.ac.jp/ai/intra_data/KawaiN/Kawai-Matsuzawa-Magical_number_5_in_a_chimpanzee.pdf.

[35] George A. Miller, *The Magical Number Seven Plus or Minus Two: Some Limits on Our Capacity for Processing Information*, 63 PSYCHOL. REV., no. 2, 1956, at 81.

[36] Cowan, *supra* note 34.

more fluent in a second language, what was previously a stream of individual sounds can subsequently be chunked into meaningful units that exist at the word level, and even the phrase level. The larger the chunks individuals can process and reconstruct, the greater their ability develops to repeat longer and more complex sentences.

Given these parameters, it can be assumed that nonnative language speakers' proficiency in their second language affects their working-memory capacity in that language; novice language learners hold fewer items in working memory than advanced learners.[37] As second language learners' proficiency in the new language advances and they become more fluent, their working memory capacity advances as well. Furthermore, the more proficient second language learners become, the more likely they are to be able to chunk the language into meaningful units, thus improving their ability to repeat phrases.[38] In repeating the EOR utterance, examinees would need to deconstruct what they heard by accessing long-term memory and by processing the sentence into meaningful chunks of information. They then must reconstruct the chunks in order to reproduce the sentence. The more proficient nonnative speakers are with the new language, the more accurate they can be expected to be at repeating a phrase they hear in that language.

Evaluation of EOR Tests

Applying the evaluative criteria discussed earlier, it is possible to measure the potential for using EOR to test speaking ability and to discover EOR strengths that are not present in the OPI. This approach also reveals certain weaknesses inherent in the EOR approach. First, reliability can be established, as it is possible to consistently administer independent items to all examinees.[39] The use of EOR in a speaking assessment may improve test/retest reliability because it can target specific grammar and vocabulary that examinees might not utter spontaneously.[40] In addition, multiple instances of the same grammatical structures or words can be used so raters can determine whether the examinees consistently perform the task. Using EOR would not eliminate the need for raters to score the recorded responses, but as the responses are narrowly defined, raters would not require as much training to be able to score the utterances.

In terms of validity, using EOR may have some benefits. Given that the EOR can be programmed to prompt examinees to say several specific phrases in a short time frame, this technique will produce many more independent samples, which will improve content coverage. In addition, the ex-

---

[37] Mary Lee Scott, *Auditory Memory and Perception in Younger and Older Adult Second Language Learners*, 16 STUD. IN SECOND LANGUAGE ACQUISITION, no. 3, 1994, at 263.

[38] Maurits W.M.L. van den Noort, Peggy Bosch & Kenneth Hugdahl, *Foreign Language Proficiency and Working Memory Capacity*, 11 EUROPEAN PSYCHOL., no. 4, 2006, at 289.

[39] CHRISTINE COOMBE, KEITH FOLSE & NANCY HUBLEY, A PRACTICAL GUIDE TO ASSESSING ENGLISH LANGUAGE LEARNERS (2007).

[40] G. Henning, *supra* note 22.

aminees cannot avoid unfamiliar vocabulary or grammatical constructs they are unable to recognize, process, and repeat. For example, if test creators wanted to determine if examinees had acquired past, unreal conditional clauses without *if*, they could present the prompt, "*Had I not gone to the store yesterday, I would not have been able to buy the toy*." This sentence would target a structure that does not occur frequently in spontaneous speech, yet might be important for testing purposes. An examinee reconstruction of, "*If I did not go to the store yesterday, I would not buy the toy*," would be evidence of some language deficiency. In this example, the meaning of the sentence has changed, indicating that the examinee was unable to comprehend the sentence. In an OPI-type interview, the examinee might not be prompted to use this type of grammar, or might construct a response that avoids difficult sentence structure. Using EOR can increase content validity because a wide range of topics, vocabulary, and structures can be sampled. However, since EOR is an indirect test of speaking, it would have weaker construct validity for testing conversational skills, as successfully repeating a sentence in a controlled environment might not indicate that the structure would be reproduced in natural speech.[41] A test using EOR would not reveal whether the individuals know *when* to use a specific grammatical structure, only *whether* they are capable of doing so.

The greatest benefit of using EOR testing, though, might be practicality. EOR is relatively inexpensive to administer and rate.[42] If the purpose of the assessment is to quickly determine if an interpreter is needed, EOR could be a viable, reliable, and practical way to get a basic assessment of speaking ability. The results of an EOR test can easily be graded by a human, but since the language produced by an EOR subject is narrowly defined, it might be even more practical if the rating can be determined using automatic speech recognition (ASR), or a combination of both, to score the assessment. In practice, using a combination of EOR with ASR would have a nonnative English speaker repeat sentences prompted by a computer. The computer would record and process the speaker's responses using speech recognition software, then run the responses through a series of software programs that quickly produce a "score" that acts as an objective measurement of language proficiency.

V. Validation of Elicited Oral Response Testing in Academic Settings

To determine the degree to which ASR-scored EOR testing could predict speaking ability, an EOR test was administered to students in an intensive English program (IEP) associated with a large university, in conjunction

---

[41] Rosemary Erlam, *Elicited Imitation as a Measure of L2 Implicit Knowledge: An Empirical Validation Study*, 27 Applied Linguistics, no. 3, 2006, at 464.

[42] *Id.*

with a battery of additional placement tests.[43] This study focused on students enrolled in an intensive English program in order to study English in preparation for university study. Participants were 179 students from various countries around the world, speaking seventeen different languages.

The study consisted of six instruments: (1) an EOR test that was scored using ASR software; (2) a placement interview with trained interviewers which assessed speaking; (3) a writing test; and three computer adaptive placement (CAPE) exams: (4) assessing listening, (5) reading, and (6) grammar. The speaking-ability variable was explored using both the oral placement interview and the EOR test.

Results indicated that the EOR test had an internal reliability of .941 as measured by a Cronbach Alpha calculation.[44] To test the degree to which the ASR-scored EOR test results could be used to predict speaking level results, a simple regression was run. This regression was found to be significant at the á=.05 level, $F(1, 174) = 154.74$, $p<.001$, adjusted $r^2=.468$, indicating that about 47% of the variance in the speaking score levels could be explained by the results of the ASR-scored EOR test results (see Figure 1).

The evidence suggests that ASR-scored EOR tests could be used to predict speaking ability, especially in making decisions such as placement testing. It also seems to show great potential as a cost-effective alternative to conducting expensive face-to-face speaking-proficiency interviews.

The strengths of the EOR test were found in reliability, validity, and practicality. Reliability was increased as multiple samples of the same objective were tested. Reliability in rating was also improved, since it was easier to consistently score items. The EOR approach allowed test designers to sample a wide range of speech structures that required examinees to respond to items they might otherwise have avoided, thus implying an improvement in content validity.[45] The greatest advantage of ASR-scored EOR tests was found in the practicality of the approach.[46] In this instance, scoring was done automatically via technology since the infrastructure was in place,[47] but even if ASR results were rated by hand, no specialized training to administer and rate the test would be required.[48]

VI. EOR Case Study 1: Hamza

During Summer 2010, an EOR test was used for an actual case involving an immigrant to the United States.[49] Hamza came from an African na-

[43] Troy Cox & Randall Davies, *Using Automatic Speech Recognition Technology with Response Testing*, CALICO Journal, Sept. 2012 at 601.

[44] Jennifer Larson-Hall, A Guide to Doing Statistics in Second Language Research Using SPSS (2010).

[45] Cox & Davies, *supra* note 43.

[46] *Id.*

[47] *Id.*

[48] *Id.*

[49] As the case did not proceed, no case reference is available.

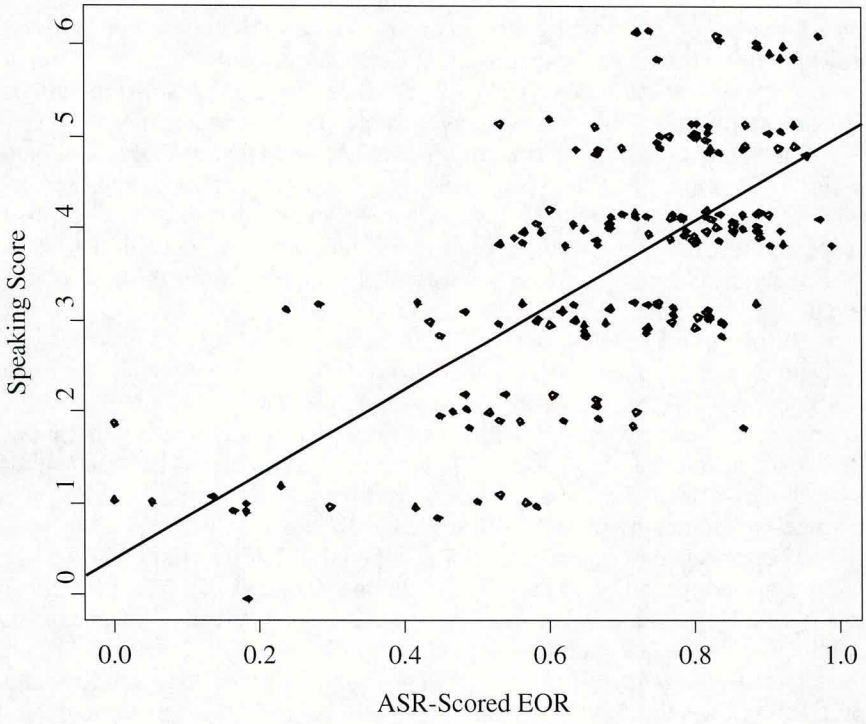

FIGURE 1. RELATIONSHIP BETWEEN ASR-SCORED EOR TEST AND SPEAKING SCORE LEVEL

tion, where French is the official language along with thirty regional, indigenous languages. As previously mentioned, Hamza, who spoke English as a fourth language, had been arrested and charged with sexual battery on an airplane. The defense counsel from the local federal defender's office approached the authors, Eggington and Cox, with two questions: 1) Did Hamza have sufficient English language proficiency to understand that he had waived his Miranda rights?; and 2) Did he have sufficient language proficiency to confess what the state claimed he had confessed? This presented an ideal opportunity to determine how well the EOR test would function in relation to the other language assessments administered to Hamza.

The research methodology included the following:

1) An initial oral proficiency interview with a certified OPI tester;
2) A second oral proficiency interview with an experienced forensic linguist;
3) Two independent analyses of a videotaped interview with the detectives who had arrested Hamza; and
4) An EOR test as described above.

All the meetings with Hamza were audio-recorded so the data could be preserved and verified. Prior to administering any of the tests, and between the various assessments, Hamza was put at ease with friendly, non-threatening conversation. For example, for the initial oral proficiency interview, the interviewer—who had studied French—spoke with Hamza in French for a short while and would ask Hamza how to say certain words in that language. The formal assessments did not begin until Hamza's body language indicated that he was relaxed. Furthermore, Hamza was not told at any time why his oral language ability was being measured. This was done to make sure he would not feel pressured to under-perform and misrepresent his language ability.

A. Findings

1. Initial OPI

In the initial interview conducted, Hamza provided evidence of being a Novice-High speaker of English. While he could at times appear fluent, especially with familiar topics, the expert believed that this was mostly due to his acquisition of learned material and stock phrases. Whenever the topical domain was unfamiliar to him, his English language ability was so low that the expert, Cox, could not assume he had comprehended what was occurring during his interrogations with police.

A number of examples from the oral interview confirmed this assumption:

1) When asked about his employment, he seemed to indicate that he was a security guard in a food court at a mall. The interviewer restated the information and Hamza nodded his head and gave verbal indication confirming that to be the case. Later, however, he stated that he was a guard in a drugstore.
2) When asked if he was married and had children, he replied that he was, but that his family was in Africa. Later, when asked about his wife, he stated that she was in California. This inconsistency demonstrated that he had not understood the initial question, although at the time of his original response, he had appeared to.
3) He did not understand what "hobbies" are until multiple examples had been given.
4) He did not understand his assigned task to "ask the interviewer some questions" until he was given an example of how to begin. When he did ask questions, they were ill formed. For example, he asked, "How many kids you have?" and, "Are your daddy alive?"
5) The official oral interview included a statement about the procedure. One of the last sentences in the statement was, "At some point during the interview I may ask you to participate in a role-play situation with me. I will introduce the role play in English, then you and I will act out the situation in English." At the end of

> the statement, Hamza was asked if he had any questions, and he replied that he did not. He was then asked if he knew what a role play was and he replied that he did not, thus illustrating that even when given the opportunity for clarification of the process, rather than indicating he did not understand, he feigned comprehension.

After conducting this interview, Cox concluded that there was sufficient evidence to indicate that Hamza needed an interpreter in order to comprehend anything more than basic survival language such as providing a name and asking for directions.

2. Second OPI

The second OPI, which was administered a number of weeks later by an experienced forensic linguist and language tester, confirmed that Hamza's language ability was too low for him to converse competently without an interpreter.

3. Interrogation Analysis

Throughout the initial interrogation with the police officers, Hamza seemed to agree to nearly every question that was presented to him. For example, the detectives asked if he was familiar with his Miranda rights and whether he had seen portrayals of police "reading the rights" on television shows, to which he nodded. There were no follow-up questions, though, to determine whether he truly understood what those rights were. The most incriminating assertion transpired when Hamza was asked if he touched the passenger "on the butt." He responded verbally in the affirmative, yet with his hand he was pointing to his back, not his bottom.

4. Elicited Oral Response Test

The EOR test was administered using a software program and a laptop computer. Prior to administration of the test, Hamza was given a verbal description of the test and the directions. The testing program was then administered to him. The first few screens contained directions with an opportunity to test out the headset and microphone to make sure the prompts were audible and the microphone would record. The program then emitted a sentence, so as to provide Hamza with an opportunity to practice before the assessment began. The test administrator asked if Hamza had heard the sentence. Hamza indicated he had, but he then proceeded to read aloud the directions on the screen indicating that he had not understood the examiner's question. After the test had begun, it became evident that the audio was not working and the test had to be restarted. Thus, although Hamza had not asked for help and had indicated that everything was okay, his language ability was apparently too low for him even to follow the simple directions for the test.

When Hamza finally started the test, he would listen to a sentence and mumble out loud, but he only attempted a few sentences. When the test was finally rated, it was discovered that, of the sixty sentences, Hamza attempted to repeat just four of them. Of the four, he repeated only one entirely correctly. In the interest of test security, specific examples of Hamza's responses cannot be provided.

B. Discussion of Findings

The prosecutor dropped the case against Hamza before it went to trial, on the basis that Hamza's alleged confession was due to his poor language ability. In addition, further investigations revealed that the alleged victim's claims were not reliable. Every language assessment examined by the experts indicated that Hamza did not have the language ability to give the police and court accurate information. The EOR test was no exception. The benefit of EOR, however, was that it was administered at a fraction of the cost of each of the other assessments. In fact, it could have been administered by the police prior to their interrogation of Hamza and prior to their reading of the Miranda rights.

Even though the charges were dropped, there was still a high personal cost for Hamza. Due in part to his low socioeconomic status, he was unable to pay bail, and he was consequently imprisoned for over five weeks and separated from his family until the prosecution dropped the charges. On a societal level, there was a significant financial cost that may have been avoided. Because he could not afford an attorney, the state had to pay for both his prosecution and his defense. In addition to the expenses incurred for attorneys and for housing Hamza as an inmate, the state had to pay language experts to conduct interviews, analyze the interrogation, write up reports, and so on. Had the police conducted a simple prescreening language assessment prior to their questioning of Hamza, the overall costs may have been reduced to the interpreter's fees and the cost of housing Hamza until a competent interpreter had been found.

While a single descriptive case study of the effectiveness of EOR for legal contexts cannot be generalized, this study should raise awareness of an important issue: the heavy cost of misclassification error. Based upon Hamza's experience as well as the extensive twenty-year experience of one of the authors, it appears to be standard practice that when a suspect or other person of interest interacts with police or the court system, police or court officials classify the individual at the initial point of interaction as either in need of an interpreter or having sufficient basic language ability to interact without an interpreter. In essence, the officials in such circumstances perform a language assessment, though in large part it is an informal assessment administered by someone who is untrained. This practice may be simpler than administering a high-standard test, but the lack of training and informal nature of the interaction adversely affects the validity and reliability of the assessment.

If the police underestimate a suspect's fluency, they can be said to have made a false negative classification error. This error is not problematic to the suspect because the error would result only in the individual being provided with an interpreter even though she might not need one. While it might be redundant to use an interpreter when an individual's English language ability is sufficiently developed, it still would not harm that individual; rather, it would give the individual interacting with the court system an additional resource.

However, if the police overestimate a suspect's fluency, they can be said to have made a false *positive* classification error. False positive errors are potentially detrimental because the police assume that the suspect has a higher level of proficiency than she actually does, resulting in a significantly higher likelihood of misinformation. Implementing a speaking proficiency measurement that could be administered by untrained personnel would be a cost-effective alternative to the status quo. An interpreter would then be consulted whenever 1) the police suspected that the subject's English language ability was too low; 2) the test result indicated it was too low; or 3) when both the police's informal interview and the test result indicated it was too low. An interview would only proceed without an interpreter if both the police and the test indicated that the person had the ability to fully comprehend English. The EOR test described above would work well for this scenario because, as discussed above, it is easily administered, non-experts can reliably grade it, and it provides timely results.

This option is even more attractive given that the EOR test can be rated automatically, as will be discussed in the following case study. As technology enables increasing computer mobility, the potential of having the program available on a smart phone or other mobile technology would further increase the practicality of administering this test.

The value of this case study, though, is the demonstration that someone who could appear fluent in English at times, was, in reality, a Novice-High speaker who could not understand what he was agreeing to, and that the necessary determination of an individual's language proficiency can be made using EOR as a valid, reliable, and practical testing procedure.

VII. EOR Case Study 2: *Escamilla v. Cuello* and Cabrera[50]

The authors of this article were recently involved in a case that generated considerable national and international press. The case, *Escamilla v. Cuello*, moved from the Arizona Superior Court to the Arizona Supreme Court and provided another opportunity to evaluate EOR with ASR in legal contexts.[51] By way of providing background information, the City of San Luis is located on the north side of the U.S.-Mexico border near Yuma, Arizona. Twenty-five thousand people live in San Luis, while approximately

[50] Escamilla v. Cuello, 282 P.3d 403 (Ariz. 2012).
[51] *Id.*

250,000 people live in San Luis Río Colorado, a parallel Mexican city on the other side of the border. Almost everyone in San Luis is fluent in Spanish. English proficiency ranges from minimal to highly fluent. For example, almost all of the city's elected leadership is fluent in both English and Spanish. San Luis' government is fragmented with various factions competing for the mayor's office and city council seats. Attempted recalls and other political dramas occur often.[52] In fact, the city had six different mayors from 1996 to 2006.[53] Alejandrina Cabrera belonged to a faction that opposed the current mayor,[54] and her faction had attempted to recall the mayor twice. Along with others in her faction, Cabrera placed her name on the ballot for the 2012 city council election.[55]

The current mayor knew that Cabrera did not speak English well. He charged the city attorney to challenge her eligibility based upon various Arizona laws that require English proficiency for elected officials and all government business be conducted in English. For example, the foundational Arizona statehood 1910 Enabling Act states:

> That said state shall never enact any law restricting or abridging the right of suffrage on account of race, color, or previous condition of servitude, and that ability to read, write, speak, and understand the English language sufficiently well to conduct the duties of the office without the aid of an interpreter shall be a necessary qualification for all state officers and members of the state legislature.[56]

The city attorney decided he needed someone to test Cabrera's English proficiency and determine if she could function as a member of the city council in English without the aid of an interpreter. After conducting an online search, he found the name of one of the authors of this article, William G. Eggington, who had been involved in a number of court cases over the years that required language testing of nonnative English speakers. The city attorney contacted Eggington and, after independently determining that there appeared to be no racist motivations behind the legal challenge, Eggington chose to offer his services.

Cabrera was born in Yuma, Arizona, but grew up on the Mexican side of the border. All of her schooling was in Spanish until she was about seventeen or eighteen, at which point she moved to San Luis and graduated from a San Luis high school's bilingual program. For a brief time, she attended college in Mexico. She currently lives in San Luis where she functions almost totally in Spanish.

---

[52] Richard Ruelas, *Politics in San Luis are Personal*, ARIZ. REPUBLIC, May 19, 2012, *available at* http://www.azcentral.com/arizonarepublic/arizonaliving/articles/20120519san-luis-politics-personal.html.

[53] *Id.*

[54] *Id.*

[55] *Id.*

[56] ARIZ. REV. STAT. ANN. § 20.

An evidentiary hearing was scheduled for January 13, 2012, to determine whether an expert should test Cabrera. The following extract from the Yuma Sun, a local newspaper, describes what happened at the hearing:

> Earlier in the hearing, and before the judge's decision, attorney John Minore, who represents Cabrera, told the court he would be willing to put his client on the stand to answer questions and read from some documents. But, he said, the matter was politically motivated, based on two unsuccessful campaigns to recall the San Luis mayor, and argued that while Eggington's test may be used to determine whether someone is proficient in English, there is no established standard of English proficiency.
>
> However, when Cabrera was called to the stand, she was unable to answer a question from Minore asking her which high school she graduated from. Although she was able to give replies to questions asking her name and where she was born, she could not answer the graduation question despite it being asked three times.
>
> After her third failed attempt to answer the question, Judge Nelson dismissed her from the stand and issued his ruling.[57]

The judge issued a ruling that Eggington should test Cabrera's English proficiency. Eggington flew to Yuma on January 16, 2012 and tested Cabrera's proficiency the next day. He administered an OPI-based protocol, as well as a fifteen-minute computer-based EOR test. The laptop computer used in the test was connected to the Internet, and transmitted, in real time, Cabrera's responses to the ASR program at Brigham Young University, where the EOR was scored using ASR software.

Results from the OPI indicated that Cabrera's proficiency was in the Novice-High to Lower-Intermediate range.[58] Troy Cox, the other author of this paper, conducted the standard independent rater reliability review of the test's audio recording and rated her as Novice-High.[59] The EOR test, using ASR as the objective scoring method, rated her as Minimally-Proficient. It should be noted that the EOR test results were sent via text message to Eggington's iPhone within twenty minutes of the conclusion of the test, thus demonstrating the feasibility of using the EOR with ASR as an objective means of determining language proficiency while providing real-time scoring from a remote location.

Based upon these analyses, Eggington expressed his opinion that Cabrera would be unable to understand council-meeting discussions or comprehend council reading material without the aid of interpreters and transla-

---

[57] James Gilbert, *Candidate's English Fluency to be Further Tested,* YUMA SUN, Jan. 13, 2012, *available at* http://www.yumasun.com/articles/city-75937-cabrera-san.html#ixzz1l2p4ObE6.

[58] ACTFL PROFICIENCY GUIDELINES 2012, *supra* note 7.

[59] *Id.*

tors. On January 25, 2012, Eggington appeared in court and underwent a one-hour examination followed by a three-hour cross-examination.

At 8:30 p.m. that same day, the presiding judge concluded that Cabrera was ineligible to stand for election and that her name should be stricken from the ballot.[60] Subsequently, Cabrera appealed the case.[61] On February 7, 2012, the Arizona Supreme Court upheld the lower court's decision.[62] The following extract from the supreme court's published opinion indicates the admissibility of Eggington's language testing protocol, including using EOR with ASR pursuant to Arizona evidentiary rules:

> Arizona Rule of Evidence 702 provides the requirements for admitting expert testimony. Dr. Eggington's curriculum vitae shows his extensive expertise in linguistics. To determine the language skills necessary to hold the office of city councilmember, he reviewed a random sampling of San Luis City Council meeting minutes, agendas, and reports, plus audio recordings of council meetings for a two-year period. He also had Cabrera perform three proficiency tests, two of which are widely used by government agencies to determine language proficiency and a third that has been published in peer-reviewed articles. His opinion that Cabrera "has minimal survival proficiency" and "could not adequately function as a Council member in the Council meetings" was based on these tests, his interviews of her, and his review of the city council materials. Rule 702's requirements were met.[63]

Regardless of the socio-political merits of the case, the outcome demonstrates that a combination of the Elicited Oral Response testing protocol using Automatic Speech Recognition software provides a valid, reliable, and practical solution for the need to provide accurate language assessment.

Conclusion

As immigration and linguistic diversity continues to increase, it is important for our society's legal system to be aware of the communicative problems faced when interacting with nonnative English speakers, as well as to develop a valid, reliable, and practical means of determining a subject's oral proficiency. As this article has shown, EOR testing with ASR scoring can be considered as a viable option.

Thus far in this paper, our discussion of the effectiveness of using EOR testing with ASR scoring has been somewhat clinical, as dictated by the expectations of a law review article. We beg your forbearance as we step outside these expectations and enter the social commentary domain.

---

[60] Ruelas, *supra* note 52.
[61] *Id.*
[62] *Id.*
[63] Escamilla, 282 P.3d at 407.

As noted at the commencement of this article, the majority of limited English speakers in the U.S. are from Spanish-speaking backgrounds. As such, this segment of the total U.S. population has sufficient population numbers and, in many locations, sufficient population density to form a strong group identity. When a large proportion of a social group begins to experience actual or perceptual injustice, a shared cultural attitude can develop that undermines attempts at developing social cohesion. This mistrust of the "system" is akin to that expressed by Langston Hughes with respect to how Blacks have been historically treated by the law:

> That justice is a blind goddess
> Is a thing to which we blacks are wise
> Her bandage hides two festering sores
> That once perhaps were eyes[64]

In this article, we have discussed potential areas of the U.S. legal system that are vulnerable to accusations of injustice regarding interactions with limited English speakers, in particular with the largest group of limited English speakers: Spanish speakers. It is our hope that the application of Elicited Oral Response testing using Automatic Speech Recognition can help avoid festering sores of injustice that can contribute to long-term social problems.

[64] KEVIN BOYLE, ARC OF JUSTICE: A SAGA OF RACE, CIVIL RIGHTS, AND MURDER IN THE JAZZ AGE (2004) (quoting Langston Hughes).